DISK ARRAY DEVICE HAVING SPARE DISK DRIVE AND DATA SPARING METHOD

CROSS REFERENCE TO RELATED APPLICATION

The basic foreign application filed on October 15, 2003, No. 2003-354557 in Japan is hereby incorporated by reference.

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention relates to a disk array device having plural disk drives and spare disk drives provided with the RAID function, in which data is copied from the plural disk drives to one or plural spare disk drives in parallel, and a control method for the same.

Description of the Related Art

Conventionally, a disk array device is provided with spare disk drives in addition to the RAID function to improve reliability. In such a disk array device, when a trouble has occurred in one of disk drives constituting the RAID, data is recovered from the other disk drives constituting the RAID and copied to the spare disk drives (correction copy), whereby the disk array device is recovered from a degenerate operation of the RAID to a normal access state. However, since the disk array device enters the degenerate operation at the time of data recovery, there are problems in reliability and performance

thereof. Thus, the reliability is improved by analyzing an error history of disk drives to predict the likelihood of a failure and copying data in disk drives with the high likelihood of a failure to spare disk drives before a failure occurs (dynamic sparing).

However, recently, a disk drive tends to have a larger capacity and is often used in a high load environment. Thus, time required for the above-mentioned correction copy and dynamic sparing is increased. Therefore, there is higher likelihood of occurrence of a multiple disk failure as described below. For example, it is likely that a trouble frequently occurs in a copy source disk drive and a READ operation cannot be continued during a copying operation for the dynamic sparing. In this case, the disk drive is blocked, and a data recovery operation (correction copy) is carried out for a parity group (i.e., ECC (error correcting code) group) to which the disk drive belongs. However, it is likely that a trouble such as partial sector fault or hard disk blocking occurs in another disk drive in the identical parity group during this correction copy operation. If such multiple disk failure occurs, data is lost.

SUMMARY OF THE INVENTION

It is an object of the present invention to reduce the risk of data lost due to a multiple failure of disks in a disk

array device of the RAID constitution having spare disks.

It is another object of the present invention to provide a disk array device which is capable of performing high-speed dynamic sparing even at the time of high load.

A disk array device, which can be connected to a host device so as to be capable of communicating with each other, in accordance with one aspect of the present invention is a disk array including: a disk array control unit which performs control of the entire disk array device; a host side data transfer control unit which controls data transfer to and from the host device; a disk array including at least plural data disk drives which constitute one parity group and one or more spare disk drives, wherein the one parity group has a large number of data stripes which are formed over storage areas of the plural data disk drives and the large number of data stripes can be partitioned into two or more sets of the data stripes; a cache memory which is used for temporary storage of data to be transferred between the host device and the disk array; and a subordinate side transfer control unit which controls data transfer to and from the disk array. Further, the disk array control unit has a prediction section which predicts the likelihood of occurrence of a failure for each data disk drive and a divided data copy section which selects two or more data disk drives out of the plural data disk drives as objects of divided data copy according to the predicted likelihood

occurrence of a failure, selects two or more divided storage areas by selecting one divided storage area from each of the selected two or more data disk drives, the selected two or more divided storage areas belonging to different sets of the data strips in the parity group, and controls the subordinate side transfer control unit and the cache memory so as to copy data in the selected two or more divided storage areas to the one or more spare disk drives.

In one embodiment, the disk array control unit further includes a dynamic sparing section which selects one data disk drive as an object of dynamic sparing out of the plural data disk drives according to the predicted likelihood of occurrence of a failure, selects remaining divided storage areas, from which data has not been copied by the divided data copy section, from the selected one data disk drive, and controls the subordinate side transfer control units and the cache memory so as to copy data in the selected remaining divided storage areas to the spare disk drives.

In one embodiment, in the case in which the predicted likelihood of occurrence of a failure of a first data disk drive in the plural data disk drives has reached a first level, the divided data copy section selects at least the first data disk drive and a separate second data disk drive as objects of the divided data copy. In the case in which the predicted likelihood of occurrence of a failure of the first data disk drive has

reached a second level which is higher than the first level, the dynamic sparing section selects the first data disk drive as an object of the dynamic sparing.

In one embodiment, in the case in which the predicted likelihood of occurrence of a failure of the first disk drive among the plural data disk drives has reached a predetermined level, the divided data copy section selects the first data disk drive and a second data disk drive, which has the largest predicted likelihood of occurrence of a failure after the first data disk drive, as objects of the divided data copy.

In one embodiment, the divided data copy section controls the subordinate side transfer control unit and the cache memory so as to read out data simultaneously from the selected two or more divided storage areas in a course of copying the data in the selected two or more divided storage areas.

In one embodiment, the disk array control unit further includes a spare data updating section which, in the case in which a new request for writing data in the selected two or more divided storage areas from the host device is received from the host side data transfer unit after the data in the selected two or more divided storage areas is started to be copied to the one or more spare disk drives, controls the subordinate side transfer control unit and the cache memory so as to write the new data received from the host device in the selected two or more divided storage areas and write the

data in the spare disk drives as well.

In one embodiment, the divided data copy section selects first and second data disk drives as objects of the divided data copy from the plural data disk drives, selects a first divided storage area belonging to a set of front side data stripes in the parity groups from the first data disk drive, selects a second divided storage area belonging to a set of rear side data stripes following the set of front side data stripes from the second data disk drive, and controls the subordinate side transfer control unit and the cache memory so as to copy data in the first and the second divided storage areas to the spare disk drives.

In one embodiment, the two or more divided storage areas selected from the two or more data disk drives have substantially the same size.

In one embodiment, the two or more divided storage areas selected from the two or more data disk drives have different sizes according to the predicted likelihood of occurrence of a failure of the two or more data disk drives.

In one embodiment, the prediction section stores an error occurrence history for each of the data disk drives and predicts the likelihood of occurrence of a failure for each of the data disk drives on the basis of the stored error occurrence history.

A data sparing method in accordance with another aspect of the present invention is applied to a disk array device,

which can be connected to a host device so as to be capable of communicating with each other, including: a disk array control unit which performs control of the entire disk array device; a host side data transfer control unit which controls data transfer to and from the host device; a disk array including at least plural data disk drives which constitute one parity group and one or more spare disk drives, wherein the one parity group has a large number of data stripes which are formed over storage areas of the plural data disk drives and the large number of data stripes can be partitioned into two or more sets of the data stripes; a cache memory which is used for temporary storage of data to be transferred between the host device and the disk array; and a subordinate side transfer control unit which controls data transfer to and from the disk array. This method includes: a step of predicting the likelihood of occurrence of a failure for each of the disk drives; a step of selecting two or more data disk drives as objects of divided data copy out of the plural data disk drives according to the predicted likelihood of occurrence of a failure; a step of selecting two or more divided storage areas by selecting one divided storage area from each of the selected two or more data disk drives, wherein the selected two or more divided storage areas belong to different sets of data stripes in the parity group; and a step of performing the divided data copy by controlling the subordinate side transfer control unit and the

cache memory so as to copy data of the selected two or more divided storage areas to the one or more spare disk drives.

One embodiment of this method further includes: a step of selecting one data disk drive as an object of dynamic sparing out of the plural data disk drives according to the predicted likelihood of occurrence of a failure; and a step of selecting remaining divided storage areas, from which data has not been copied to the spare disks, from the selected one data disk drive, and controlling the subordinate side transfer control units and the cache memory so as to copy data in the selected remaining divided storage areas to the spare disk drives.

In one embodiment, the step of selecting two or more data disk drives is performed in the case in which the predicted likelihood of occurrence of a failure of a first data disk drive among the plural data disk drives has reached a first level, and then at least the first data disk drive and a separate second data disk drive are selected as objects of the divided data copy. In addition, the step of selecting one data disk drive is performed in the case in which the predicted likelihood of occurrence of a failure of the first data disk drive has reached a second level higher than the first level, and then the first data disk drive is selected as an object of the dynamic sparing.

A data sparing control device in accordance with yet another aspect of the present invention is applied to a disk array device, which can be connected to a host device so as

to be capable of communicating with each other, including: a disk array control unit which performs control of the entire disk array device; a host side data transfer control unit which controls data transfer to and from the host device; a disk array including at least plural data disk drives which constitute one parity group and one or more spare disk drives, wherein the one parity group has a large number of data stripes which are formed over storage areas of the plural data disk drives and the large number of data stripes can be partitioned into two or more sets of the data stripes; a cache memory which is used for temporary storage of data to be transferred between the host device and the disk array; and a subordinate side transfer control unit which controls data transfer to and from the disk array. This data sparing control device includes: a prediction unit which predicts the likelihood of occurrence of a failure for each of the data disk drives; a divided area selection unit which selects two or more data disk drives as objects of divided data copy out of the plural data disk drives according to the predicted likelihood of occurrence of a failure and selects two or more divided storage areas by selecting one divided storage area from each of the selected two or more data disk drives, wherein the selected two or more divided storage areas belong to different sets of data stripes in the parity group; and a divided data copy unit which controls the subordinate side transfer control unit and the cache memory

so as to copy data of the selected two or more divided storage areas to the spare disk drives.

Other characteristics of the present invention will become evident from the following descriptions.

BRIEF DESCRIPTION OF THE DRAWINGS

In the accompanying drawings:

Fig. 1 is a block diagram showing a schematic structure of a computer system, which uses a disk array device in accordance with an embodiment of the present invention, and an internal structure of this disk array device;

Fig. 2 is a schematic diagram showing a structure of a disk drive resource information table;

Fig. 3 is a schematic diagram showing a structure of a spare disk drive resource information table;

Fig. 4 is a schematic diagram showing a structure of a disk array, which forms a parity group of a 3D+1P format, and a state of copying divided data to spare disk drives;

Fig. 5 is a flowchart showing a flow of a control operation for data protection of a disk array performed by disk array control units 203 and 213;

Fig. 6 is a flowchart showing a specific flow of divided data copy processing;

Fig. 7 is a flowchart showing a specific flow of dynamic sparing processing;

Figs. 8A to 8C are schematic diagrams showing a specific example of a control operation for data protection in the case in which two spare disk drives are prepared for four data disk drives constituting a disk array of a 3D+1P format;

Figs. 9A and 9B are schematic diagrams showing a modification of a data copy operation shown in Figs. 8A to 8C;

Figs. 10A and 10B are schematic diagrams showing another modification of the data copy operation shown in Figs. 8A to 8C;

Fig. 11 is a schematic diagram showing a modification of the divided data copy processing;

Fig. 12 is a schematic diagram showing another modification of the divided data copy processing; and

Fig. 13 is a schematic diagram showing a modification of a structure of a parity group.

DESCRIPTION OF THE PREFERRED EMBODIMENT

An embodiment of a disk array device in accordance with the present invention will be hereinafter described.

Fig. 1 shows a schematic structure of a computer system, which uses a disk array device in accordance with an embodiment of the present invention, and an internal structure of this disk array device.

As shown in Fig. 1, a disk array device 100 is connected to a host device 110 such as a host computer and includes plural

hard disk drives (drive numbers #0 to #N) 300 to 330 and one or two or more controllers 200 and 210. In addition, this disk array device 100 includes a control console 120 for performing operation and setting of the disk array device 100. The controllers 200 and 210 are connected to the control console 120 via LAN interface 201 and 211, respectively.

The controllers 200 and 210 include: host side data transfer control units 202 and 212 which control data transfer to and from the host device 110; disk array control units 203 and 213 which perform control of the entire disk array device 100; a cache memory 220 which is used for data transfer or the like of the host device 100 and the disk drives 330 to 330; and subordinate side transfer control units 204 and 214 which control data transfer of the disk drives 300 to 330, respectively. The controller 200 accesses the disk drives 300 to 330 via ports 400 to 430, respectively. The controller 210 accesses the disk drives 300 to 330 via ports 440 to 470, respectively. In the following description of this embodiment, a physical array of the plural disk drives 300 to 330 is referred to as a "disk array", and a logic volume with a redundant constitution according to the RAID technique is referred to as a "LU (logic unit)".

Next, with reference to Figs. 2 and 3, a data structure for managing a state of a disk array used by the disk array device 100 of the present invention will be described. Fig.

2 is a schematic diagram illustrating a disk drive resource information table 500. Fig. 3 is a schematic diagram illustrating a spare disk drive resource information table 600.

The disk drive resource information table 500 shown in Fig. 2 exists in each of the disk array control units 203 and 213 of the controllers 200 and 210. The disk drive resource information table 500 is a table for managing a state of use, a capacity, and the like of all the disk drives 300 to 330 and includes plural disk drive management areas 510 corresponding to all the disk drives 300 to 330, respectively. A ROW (row) number (ROW #0, ROW #1, ...) and a COL (column) number (COL #0, COL #1, ...) allocated to each disk drive management area 510 indicate a position of a disk drive, which corresponds to the disk drive management area 510, in the disk array device 100.

Each disk drive management area 510 has information such as a state of use 520, a full capacity 530, a spared capacity (or address) 540, and an error counter 550 with respect to the corresponding disk drive. Values indicating a situation of implementation and a situation of use for the corresponding disk drive, for example, four values of "uninstalled" and "installed" (which also distinguish whether the disk drive is used as a data disk drive or a spare disk drive), and "unused" and "used", are selectively entered in the state of use 520. In the case in which the corresponding disk drive is implemented,

a value representing a total storage capacity of the disk drive is entered in the full capacity 530. The spared capacity (or address) 540 is a field which is used at the time of copying data to the spare disk drives in the case in which the likelihood of occurrence of a failure of the corresponding disk drive has increased. A value indicating a capacity of data copied to the spare disk drives at a present point in the full capacity of the disk drive or an address (e.g., a logic block number, etc.) of a storage area from which the data is copied is entered in the spared capacity (or address) 540. A value indicating the number of times of occurrence of an access error in the corresponding disk drive is entered in the error counter 550. This value is counted up (or counted down) every time an access error occurs in the disk drive. The value of the error counter 550 is used as a value representing an error rate (frequency of occurrence of an error), that is, a degree of the likelihood of occurrence of a failure.

The spare disk drive resource information table 600 shown in Fig. 3 exists in each of the disk array control units 203 and 213 of the controllers 200 and 210. The spare disk drive resource information table 600 is a table for managing a state of a disk drive, which is used as a spare disk drive, in the disk drives 300 to 330. The spare disk drive resource information table 600 has plural spare disk drive information areas 610 which correspond to all the spare disk drives,

respectively.

Each spare disk drive information area 610 includes detailed information on a spare disk drive corresponding thereto such as a state of use 620, a spare ROW 630, a spare COL 640, and used area information [0] 650 to [m] 670. The state of use 620 selectively has a value indicating a state of use of the corresponding spare disk drive, for example, three values of "unused", "partially used", and "entirely used". In the case in which sparing of data of a certain data disk drive is performed, the state of use 620 is used for performing judgment of propriety of use of the spare disk drive. A ROW number and a COL number indicating a position of the corresponding spare disk drive in the disk array device 100 are entered in the spare ROW 630 and the spare COL 640. The used area information [0] 650 to [m] 670 is provided in association with m storage areas ([0] to [m]) included in the corresponding spare disk drive (m is the number of divisions of the spare disk drive, which is 2 in an example shown in Fig. 4 to be described later), respectively, and is information for managing for data recovery of which disk drives the m storage areas have been used, respectively. Each piece of used area information has information such as a flag in use 670, a start address 680, an end address 690, a copy source ROW 700, and a copy source COL 710. The flag in use 670 indicates whether or not a storage area corresponding thereto is currently used. The start address

680 and the end address 690 indicate a start address and an end address of the corresponding storage area. A ROW number and a COL number indicating a position of a data disk drive, in which data copied to the corresponding storage area was originally stored, in the disk array device 100 are entered in the copy source ROW 700 and the copy source COL 710.

The disk array control units 203 and 213 (see Fig. 1) of the disk array device 100 perform, using the disk drive resource information table 500 and the spare disk drive resource information table 600 which are described above, a control operation described below in order to grasp states of the disk drives 300 to 330 and protect data from a failure of the disk drives.

In the following description, as an example, an error rare (frequency of occurrence of an access error) of each disk drive is classified into three levels indicating the likelihood of occurrence of a failure of each disk drive. In other words, a "level 1", a "level 2", and a "level 3" indicate that a probability of occurrence of a failure is low, medium, and high, respectively. When an error rate of a certain disk drive is , for example, the level 2, the disk array control units 203 and 213 perform "divided data copy to spear disks", which complies with the principle of the present invention, for the disk drive to prepare for occurrence of a failure. Thereafter, when more errors occur in the disk drive and the error rate changes to,

for example, the level 3, the disk array control units 203 and 213 carries out dynamic sparing for the disk drive. Then, after completion of the dynamic sparing, the disk array control units 203 and 213 block the disk drive.

Control for protecting data from such a failure of a disk drive will be more specifically described blow. In the following description, a case of a disk array which is constituted so as to form a parity group of a 3D+1P format in accordance with the RAID-5 technique is described as an example. However, those skilled in the art should easily understand that the control complying with the principle of the present invention can be applied to a disk array of other structures in accordance with the RAID technique.

Fig. 4 shows an example of a structure of a disk array forming a parity group of the 3D+1P format.

In the example shown in Fig. 4, there is a disk array consisting of four data disk drives (DDD #0 to DDD #3) 800 to 830, which form one parity group of the 3D+1P format. In this data array 800 to 830, for example, two logic units LU0 and LU1 are defined. One logic unit LU0 is constituted by logic blocks Block 0 to Block 17, and the other logic unit LU1 is constituted by logic blocks Block 18 to Block X (a size of one logic block is, for example, 64K bytes). Four logic blocks arranged in one row in the horizontal direction in the figure, which are included in these four different data disk drives

800 to 830, respectively, form one data stripe of the 3D+1P format. Here, a logic block marked as P(Q-R) in each data strip means that it is a block storing parity data of data of other logic blocks Block Q to Block R in the same data stripe. For example, a first data stripe is constituted by three data logic blocks Block 0 to Block 2 and one parity logic block P(0-2). A second data stripe is constituted by three data blocks Block 3 to Block 5 and one parity block P(3-5). In this way, there are a large number of data stripes in the disk array 800 to 830 forming one parity group. Moreover, at least one spare disk drive (SDD #A) exists other than the disk array 800 to 830.

Fig. 5 shows an overall flow of control for protecting data from a failure of data disk drives by the data disk array control units 203 and 213 (see Fig. 1), which is applied to the disk array 800 to 830 with the structure illustrated in Fig. 4.

As shown in Fig. 5, the disk array control units 203 and 213 check whether or not there is a disk drive, an error rate of which has reached the level 2, in the disk array 800 to 830 (step 900). This check can be performed by finding the disk drive management area 510, a value of the error counter 550 of which has reached a predetermined threshold value corresponding to the level 2, from the disk drive resource information table 500 shown in Fig. 2. If such a disk drive

management area 510 is found, a disk drive in a position determined by a ROW number and a COL number corresponding to the disk drive management area 510 is detected as a disk drive, an error rate of which has reached the level 2. If such a disk drive is detected, the disk array control units 203 and 213 judge presence or absence of an available spare disk drive with reference to the spare disk drive resource information table 600 shown in Fig. 3 (step 910). If there is an available spare disk, the disk array control units 203 and 213 control the subordinate side transfer control units 204 and 214 and the cache memory 220 to perform divided data copy processing for the detected disk drive (step 920). A specific method of the divided data copy processing will be described later. On the other hand, if there is no available spare disk, the disk array control units 203 and 213 issue a warning informing the control console 110 of an increase in error rate of the disk drive to urge an operator to pay attention (step 950).

In addition, the disk array control units 203 and 213 check whether or not there is a disk drive, an error rate of which has reached the level 3, in the disk array 800 to 830 (step 930). This check can be performed by finding the disk drive management area 510, a value of the error counter 550 of which has reached a predetermined threshold value corresponding to the level 3, from the disk drive resource information table 500 shown in Fig. 2. If such a disk drive

management area 510 is found, a disk drive in a position determined by a ROW number and a COL number corresponding to the disk drive management area 510 is detected as a disk drive, an error rate of which has reached the level 3. If such a disk drive is detected, the disk array control units 203 and 213 control the subordinate side transfer control units 204 and 214 and the cache memory 220 to execute dynamic sparing for the detected disk drive (step 940). Here, as it is evident from the above description, usually the divided data copy of step 920 has already been performed for a disk drive which has become an object of the dynamic sparing. As it is seen from specific processing for the divided data copy to be described later, data in a half of an entire storage area in the disk drive has already been copied to spare disk drives by the divided data copy. Therefore, in the dynamic sparing, it is sufficient to copy data of the remaining half storage area to the spare disk drives. Thus, time required for the dynamic sparing is reduced by half from that in the past. After completion of the dynamic sparing, the disk drive is blocked, and work for replacing it with a normal disk drive is performed. Although not shown in Fig. 5, after completion of the replacement of the blocked disk drive, data is copied back (recovered) to the disk drive from a spare disk drive. After this data recovery is completed, the spare disk drive is set to a standby state and is used at the next opportunity of the divided copy processing

and the dynamic sparing.

Next, a specific method of the divided data copy processing of step 920 in Fig. 5 will be described.

In Fig. 4, an example of a specific form of the divided data copy processing, which is executed in the disk array 800 to 830, is shown together with the example of a structure of the disk array 800 to 830 which has already been described. An example of the divided data copy processing will be described with reference to Fig. 4.

In the disk array 800 to 830 shown in Fig. 4, for example, it is assumed that one data disk drive (DDD #0) 800 has reached the level 2. At this point, it is assumed that an error rate of the other data disk drives (DDD #1 to #3) 810 to 830 is the level 1. However, it is assumed that a value of the error counter of the data disk drive (DDD #1) 810 is the highest among the data disk drives (DDD #1 to #3) 810 to 830 of the level 1. In addition, at this point, it is assumed that an available spare disk (SDD #A) 840 exists.

In such a situation, divided data copy processing is executed for the disk drive (DDD #0) 800, which has reached the level 2, before it breaks down and the READ operation becomes impossible. In this divided data copy processing, the storage area (logic blocks Block 0 to Block 12) 850 in the upper half in the figure (front half) is selected out of all the storage areas of the disk drive (DDD #0) 800, and processing for copying

data of the selected upper half storage area 850 to an upper half storage area 880, which corresponds to the upper half storage area 850, in the spare disk drive (SDD #A) 840 is executed. In other words, the data of the upper half storage area 850 is read out to the cache memory 220 by the control of the subordinate side transfer control units 204 and 214 and written in the corresponding upper half storage area 880 of the spare disk drive (SDD #A) from the cache memory 220. In parallel with this, another disk drive (DDD #1) 810 having the highest error count value (i.e., the highest likelihood of occurrence of a failure) after this disk drive (DDD #0) 800 is selected. A recording place area (Block 16, P7, Block 21, ...) 870 of a lower half (rear half), which corresponds to a set of data strips of a lower half (rear half) following a set of data strips of an upper half (front half) to which a divided copy range (i.e., the upper half storage area 850) in the above-mentioned disk drive (DDD #0) 800 belongs, is selected from the selected disk drive (DDD #1) 810. Then, processing for copying data of the selected lower half recording place area (Block 16, P7, Block 21, ...) 870 to a half storage area 890, which corresponds to the lower half recording place area 870, of the spare disk drive (SDD #A) 840 is executed. In other words, data of the lower half recording place area 870 is read out to the cache memory 220 by control of the subordinate side transfer control units 204 and 214 and written in the corresponding half storage

area 890 of the spare disk drive (SDD #A) 840 from the cache memory 220.

In short, in the divided data copy processing, the two disk drives (DDD #0, #1) 800 and 810, which have the relatively high likelihood of occurrence of a failure, are selected from the disk array 800 to 830 forming one parity group, and the upper half (front half) and the lower half (rear half) storage areas 850 and 870, which belong to different sets of data stripes, are selected from the selected two disk drives (DDD #0, #1) 800 and 810, respectively. Then, data is simultaneously read from the selected upper half (front half) and lower half (rear half) storage areas 850 and 870 and written in the spare disk 840. At this point, the data simultaneously read from the two disk drives (DDD #0, #1) 800 and 810 are scheduled by the disk array control units 203 and 213 and written in the spare disk drive 840. Time required for data READ is reduced by half by the simultaneous data READ from the two disk drives. The reduction in copy time is particularly advantageous in a high load environment. Moreover, in the case in which the spare disk drive 840 has plural access ports, since multiple data WRITE to the spare disk drive 840 can be performed, speed of the divided data copy processing is increased.

Note that, in the case in which a WRITE request for new data is inputted to a logic block, for which copy has already been completed, in the disk drives (DDD #0, #1) 800 and 810

as objects of copy from the host device 110 or the like through the host side data transfer control units 202 and 212 during execution of the divided data copy processing, the disk array control units 203 and 213 control the subordinate side transfer control unit 204 and the cache memory 220 to write the data duplicately to the disk drive (DDD #0 or #1) 800 or 810 and the spare disk drive (SDD #A) 840 as objects of WRITE. Consequently, data in the spare disk drive (SDD #A) 840 can be always kept in an updated state and prepared for occurrence of a failure in the copy source disk drives (DDD #0, #1) 800 and 810.

Fig. 6 shows a specific flow of the divided data copy processing.

As shown in Fig. 6, when the divided data copy processing is started, two storage areas, which are objects of divided data copy, are selected from a disk drive, an error rate of which has reached the level 2, and a disk drive having the highest error rate (error count value) after this disk drive (step 960). In the example of Fig. 4, the half storage areas 850 and 860 are selected from the disk drive (DDD #0 or #1) 800 or 810, respectively. Here, it is important that the selected two storage areas 850 and 860 do not overlap in a set of data stripes to which the storage areas belong. In other words, all logic blocks selected as objects of copy belong to different sets of data stripes, respectively. Consequently, it becomes more

likely that data of as many number as possible are protected.

Thereafter, data is sequentially read in parallel from all logic blocks in copy object areas of the respective two disk drives, and the read data is written in logic blocks, which corresponds to the logic blocks in the copy object areas, of spare disk drives in accordance with a predetermined schedule (in a loop of step 970 and steps 990 to 1010 for one drive, and a loop of step 970 and steps 1020 to 1050 for the other drive).

Fig. 7 shows a specific flow of the dynamic sparing processing applied to a disk drive, an error rate of which has reached the level 3.

As shown in Fig. 7, when the dynamic sparing is started, it is confirmed by the disk array control unit 203 or 213 which storage area in a disk drive, which becomes an object of the processing, is a storage area for which sparing (copy to a spare disk) has already been completed (step 1100). This confirmation can be performed by referring to the spared capacity (or address) 540 of the disk drive management area 510 corresponding to the disk drive in the disk drive resource information table 500 shown in Fig. 2. Usually, since the divided data copy processing has already been performed to the disk drive, sparing has already been completed for a half storage area in the disk drive. Thereafter, data of a storage area, for which sparing has not been completed yet, in the disk drive is sequentially read out

to the cache memory 220 through the subordinate side transfer control unit 204 or 214 in an order of block numbers by control of the disk array control unit 203 or 213, and copied to a storage area, which corresponds to the above-mentioned storage area, of a spare disk drive from the cache memory 220 through the subordinate side transfer control unit 204 or 214 (a loop of steps 1110 to 1140).

Note that, in the case in which a disk drive as an object of the dynamic sparing breaks down and READ of data cannot be performed before the dynamic sparing is completed, the data to be read is recovered and copied to the spare disk drive by a correction copy operation using data in another disk drive constituting the same parity group as the data to be read. In the case in which the above-mentioned another disk drive also breaks down and READ of the data cannot be performed before this correction copy operation is completed, according to the related art, a sub-system relating to these broken-down disk drives cannot help but lapse into a system down state. However, in this embodiment, since the divided data copy has already been performed for the broken-down another disk drive, even if duplicate failures occur as described above, the data as an object of correction copy can be recovered using the data in the spare disk drive.

Next, a specific example of the above-mentioned control operation for data protection in the case in which two spare

disk drives are prepared for four data disk drives, which constitute a disk array of the 3D+1P format, will be described with reference to Figs. 8A to 8C.

As shown in Fig. 8A, in the four data disk drives DDD #0 to #3 constituting the disk array of 3D+1, in the case in which an error rate of the disk drive DDD #0 has reached the level 2 and error rates of the other disk drives DDD #1 to #3 are the level 1 (#2>#3>#1), the divided data copy processing for copying divided data to the spare disk drive SDD #A is executed for the disk drive DDD #0 having the error rate of the level 2 and the disk drive DDD #2 having the second highest error rate. In other words, data #0_UH of the upper half (front half) of the disk drive DDD #0 and data #2_LH of the lower half (rear half) of the disk drive DDD #2 are copied to the spare disk drive SDD #A.

Thereafter, as shown in Fig. 8B, in the case in which the error rate of the disk drive DDD #0 has increased to reach the level 3, the dynamic sparing processing for copying data to the spare disk drive SDD #A is executed for the disk drive DDD #0 before the disk drive DDD #0 becomes unreadable. In the dynamic sparing processing, it is sufficient to only copy data #0_LH of the lower half (rear half), which has not been spared, of the disk drive DDD #0 to the spare disk drive SDD #A. At this point, data #2_LH of the lower half (rear half) of the disk drive DDD #2 in the spare disk drive SDD #A is deleted.

At this point, in the case in which the error rate of the disk drive DDD #2 has reached the level 2 as well, the divided data copy is executed, and data #2_UH of the upper half (front half) of the disk drive DDD #2 and data #3_LH of the lower half (rear half) of the disk drive DDD #3 having the second highest error rate are copied to another spare disk drive SDD #B.

After completion of the dynamic sparing processing for the disk drive DDD #0, the disk drive DDD #0 is blocked and is replaced with a separate normal disk drive. The divided data copy processing disk drives DDD #2 and DDD #3 are used even after the divided data copy processing is completed as usual as long as error rates thereof do not reach the level 3. However, when WRITE of data to the disk drive DDD #2 or DDD #3 is performed, same data is duplicately written to the spare disk drive SDD #B as well as the disk drive DDD #2 or DDD #3, whereby data in the spare disk drive SDD #B is always kept updated.

After the replacement of the disk drive DDD #0 is completed, as shown in Fig. 8C, data is copied from the spare disk drive SDD #A to the disk drive DDD #0, and the data of the disk drive DDD #0 is recovered. After this recovery processing is completed, the spare disk drive SDD #A comes into the standby state. At this stage, the data of the spare disk drive SDD #B is continuously kept in the updated state and prepared for a future increase in error rate of the disk drives DDD #2 and

DDD #3.

Figs. 9A and 9B show a modification of the data copy operation shown in Figs. 8A to 8C.

In the above-mentioned data copy operation of Figs. 8A to 8C, in the divided data copy processing shown in Fig. 8A, the data #0_UH and #2_LH of the two disk drives DDD #0 and DDD #2 are copied to the identical spare disk drive SDD #A. However, as an modification, it is also possible to copy the data #0_UH and #2_LH of the two disk drives DDD #0 and DDD #2 to the different spare disk drives SDD #A and SDD #B, respectively, as shown in Fig. 9A. In this way, thereafter, in the case in which an error rate of the disk drive DDD #0 has reached the level 3 and the dynamic sparing processing is performed as shown in Fig. 9B, it is possible to copy the remaining data #0_LH of the disk drive DDD #0 to the spare disk drive SDD #A without deleting the data #2_LH of the disk drive DDD #2 which has already been spared. As a result, in the case in which an error rate of the disk drive DDD #2 has reached the level 2 and the divided data copy processing is performed, it is possible to copy only the data #3_UH of the disk drive DDD #3 having the second highest error rate without copying the spared data #2_LH of the disk drive DDD #2.

Figs. 10A and 10B show another modification of the data copy operation shown in Figs. 8A to 8C.

Operations of the divided data copy processing shown in

Fig. 10A are the same as those shown in Fig. 8A. However, in the dynamic sparing processing of the disk drive DDD #0 shown in Fig. 10B, the data #0_LH of the lower half (rear half) of the disk drive DDD #0 is copied to the spare disk drive SDD #B separate from the spare disk drive SDD #A to which the data #0_UH of the upper half (rear half) is copied. Consequently, the data #2_LH of the disk drive DDD #2, which has already been spared, does not have to be deleted. As a result, as in the example of operations of Figs. 9A and 9B, in the case in which an error rate of the disk drive DDD #2 has reached the level 2 and the divided data copy processing is performed, it is possible to copy only the data #3_UH of the disk drive DDD #3 having the second highest error rate without copying the spared data #2_LH of the disk drive DDD #2.

Fig. 11 shows a modification of the operation of the divided data copy processing shown in Fig. 4.

In the divided data copy processing shown in Fig. 4, the storage areas 805 and 807 having a substantially equal size (i.e., half size) are selected from the two disk drives DDD #0 and DDD #1, respectively, to perform sparing of those data. As a modification, as shown in Fig. 11, it is also possible to select storage areas 1200 and 1210 having unequal sizes from the two disk drives DDD #0 and DDD #1 to perform sparing of those data. In the example of Fig. 11, the storage area 1200 of the larger size is copied from the disk drive DDD #0 with

a higher error rate (likelihood of occurrence of a failure), and the storage area 1210 of the smaller size is copied from the disk drive DDD #1 with a lower error rate. According to such unequal divided data copy processing, a reduction effect of time required for the divided data copy is low compared with the equal divided data copy processing shown in Fig. 4. However, an effect that processing time can be reduced in the dynamic sparing processing to be performed later (which is highly likely to be performed for the disk drive DDD #0 with the higher error rate) can be expected so much more for that.

Fig. 12 shows another modification of the operations of the divided data copy processing shown in Fig. 4.

In the divided data copy processing shown in Fig. 4, the two disk drives DDD #0 and DDD #1 are selected as objects of processing. As a modification, as shown in Fig. 12, three or more disk drives DDD #0, DDD #1, DDD #2, and the like may be selected as objects of the divided data copy processing. In the example of Fig. 12, the three disk drives DDD #0, DDD #1, and DDD #2 are selected in order from one with the highest error rate, and three storage areas 1300, 1310, and 1320 of parity groups which do not overlap each other are selected and copied, respectively.

The embodiment of the present invention has been described. However, this embodiment is only an example for the description of the present invention, and it is not intended to limit the

scope of the present invention only to this embodiment. The present invention can be carried out in various other forms without departing from the gist thereof.

For example, in the above-mentioned embodiment, an error rate is classified into three levels, and the divided data copy is performed for a drive of the level 2 and the dynamic sparing is performed for a drive of the level 3. However, it is also possible to classify the error rate into more than three levels and, for drives of intermediate plural levels (e.g., the levels 2 and 3 among the four levels), perform the divided data copy corresponding to the levels (e.g., as the level is higher, a data size to be spared increases).

In addition, the above-mentioned embodiment is described with the disk array, which constitutes the parity group of the 3D+1P format in accordance with the "RAId-5" technique, as an example. However, it is evident that the present invention is applicable to a disk array, which constitutes a parity group of other forms in accordance with any one of the "RAID-1" to the "RAID-5". For example, even in the parity group of the 3D+1P format, there are the example shown in Fig. 13 and other variations in an arrangement of parity data. The present invention is applicable to any one of those variations.

In addition, in the above-mentioned embodiment, the disk array device controls the operations of the divided data copy, the dynamic sparing, and the like independently from a command

from an external device such as a host device. However, such control may be performed according to a command from an external device.